Microscopes and Telescopes: Trading in Black Boxes for a Lens with Multitexts, Network Depths, and Statistical Comparisons

Ada Wan
adawanwork@gmail.com

Abstract

Deep neural networks (DNNs) have typically been thought of as black boxes. Work on evaluation and interpretations has often tried to look into/through the box — by testing each model with various hyperparameter settings or via output generated by each model. In this paper, we examine the effects of network depth and show how it is also possible to look outside the box and arrive at an interpretation through a meta evaluation across multiple models. Following a setup similar to Wan (2022), we perform systematically controlled experiments in conditional language modeling with the Transformer and multiway parallel data, controlling for the number of layers in depth, holding all other hyperparameters constant. We present visualization of our results and substantiate our interpretation with statistical comparisons confirming that there are more instances of significant differences between pairs in deeper models than in shallower models. That is, all else being equal, the deeper models magnify, like microscopes, differences in raw data statistics, while the shallower models, much like telescopes, neutralize/compress them.

1 Introduction

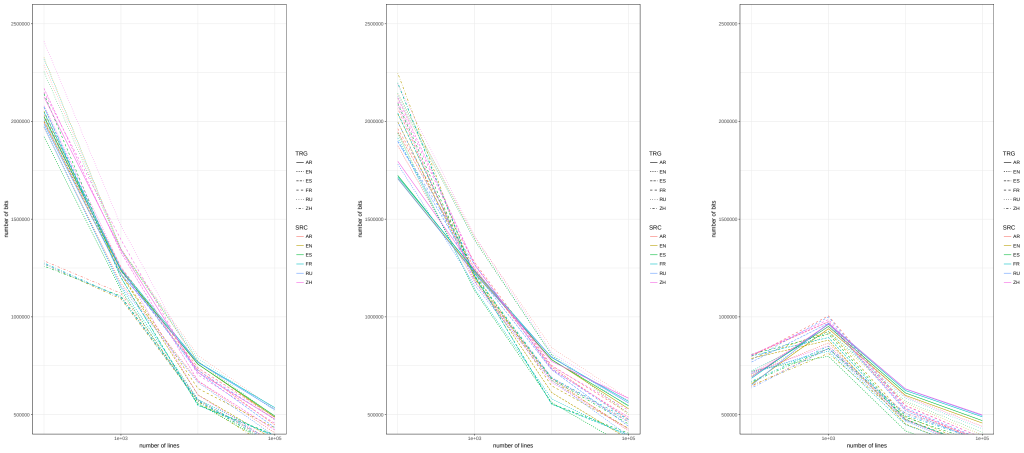
DNNs are often thought of as black boxes. Work on evaluation and interpretability mostly focuses on looking within each model, i.e. into each "box", through hyperparameter tuning. In this paper, we show how the results from sets of conditional language models (CLMs) from systematically controlled experiments can illustrate interpretive insights. We follow the meta-evaluation method from Wan (2022)[1] for the setup of our experiments. But instead of studying the disparity in results between representation granularities, we seek to understand the effects of network depth on the Transformer (Vaswani et al., 2017) using a data-centric approach with one identical hyperparameter setting for all models.

1.1 Summary of Contributions

We perform systematically controlled experiments in CLMing with parallel data to study Transformer network depth (1-6 layers) across 4 data sizes (10^2-10^5 lines) and 3 representational granularities (in character, byte, and word).

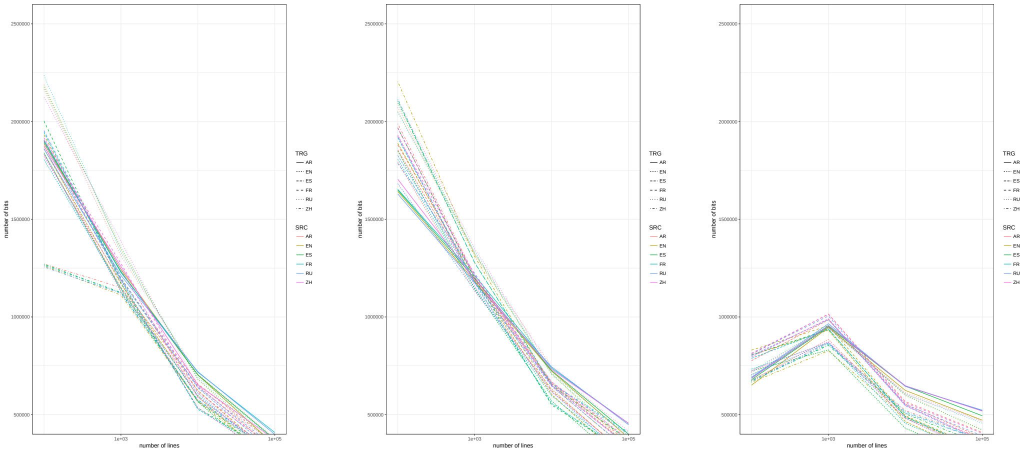
1. We report and visualize our experimental results from 6480 CLMs based on 3 runs and evaluate their differences via statistical comparisons.
2. We find deeper models magnify differences in data statistics while shallower models neutralize/compress them. (We use the analogy of "microscopes" and "telescopes" in our title to loosely capture this epiphenomenon.)
3. We observe, based on our one systematically controlled setting, that it is possible for shallower models to outperform deeper ones.

[1] Wan (2022) was a reformulation based on Wan (2021). For the purpose of this work, the two can be interchangeably referenced.



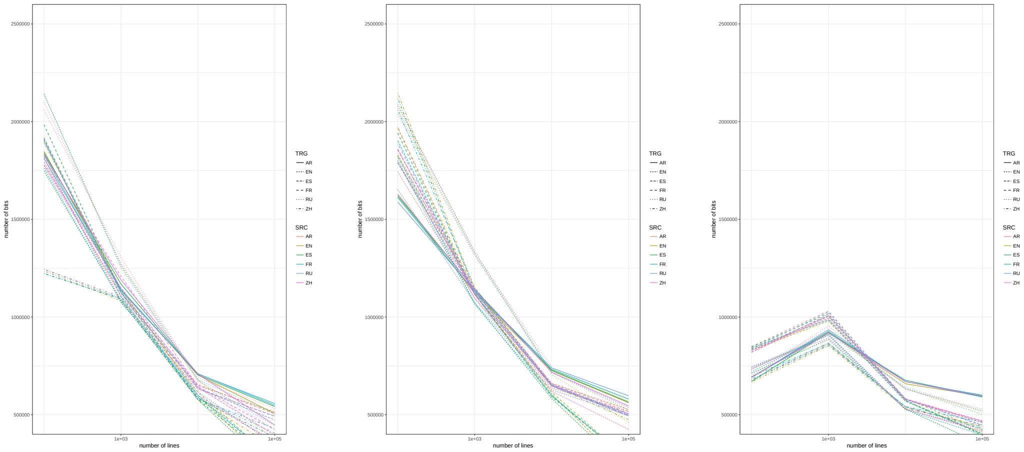
(a) CHAR (b) BYTE (c) WORD

Figure 1: 1-layer models



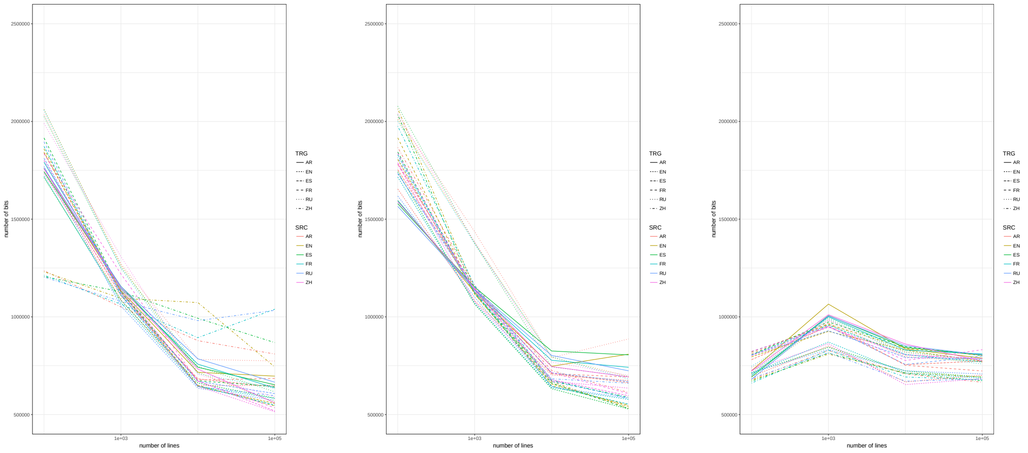
(a) CHAR (b) BYTE (c) WORD

Figure 2: 2-layer models



(a) CHAR (b) BYTE (c) WORD

Figure 3: 3-layer models

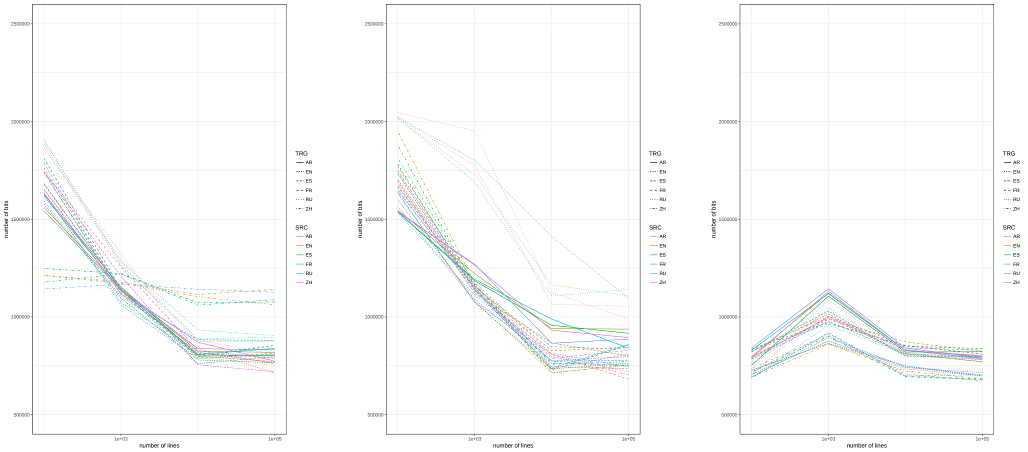


(a) CHAR

(b) BYTE

(c) WORD

Figure 4: 4-layer models

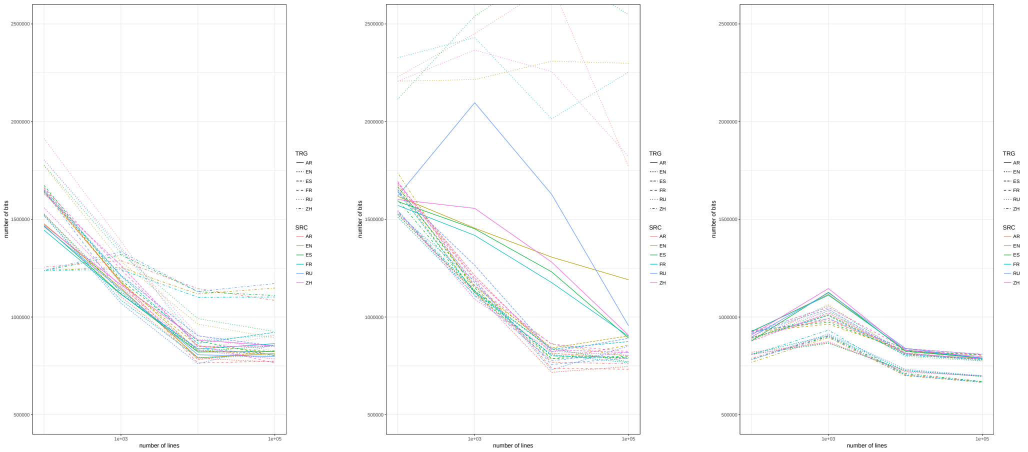


(a) CHAR

(b) BYTE

(c) WORD

Figure 5: 5-layer models

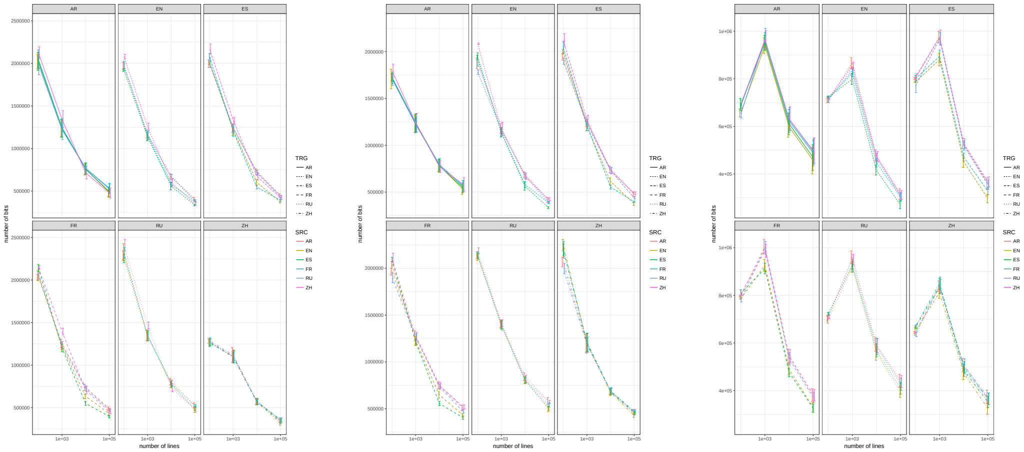


(a) CHAR

(b) BYTE

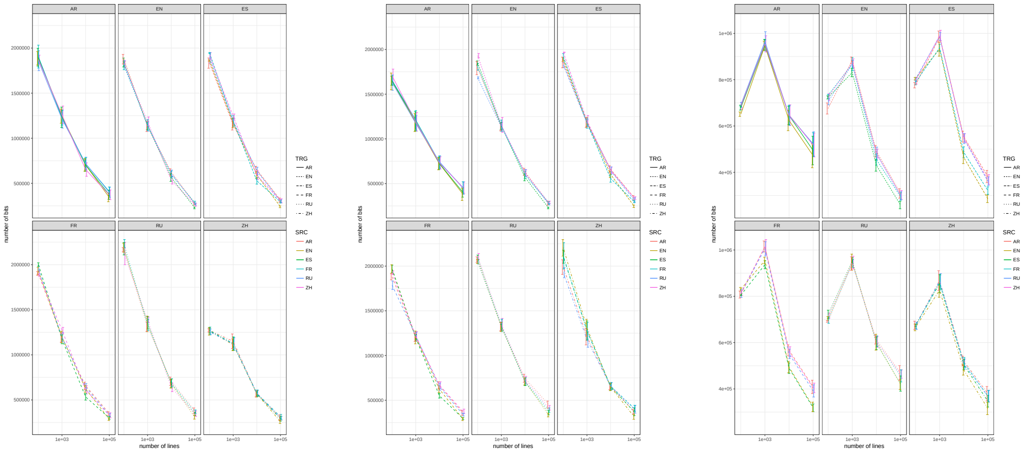
(c) WORD

Figure 6: 6-layer models



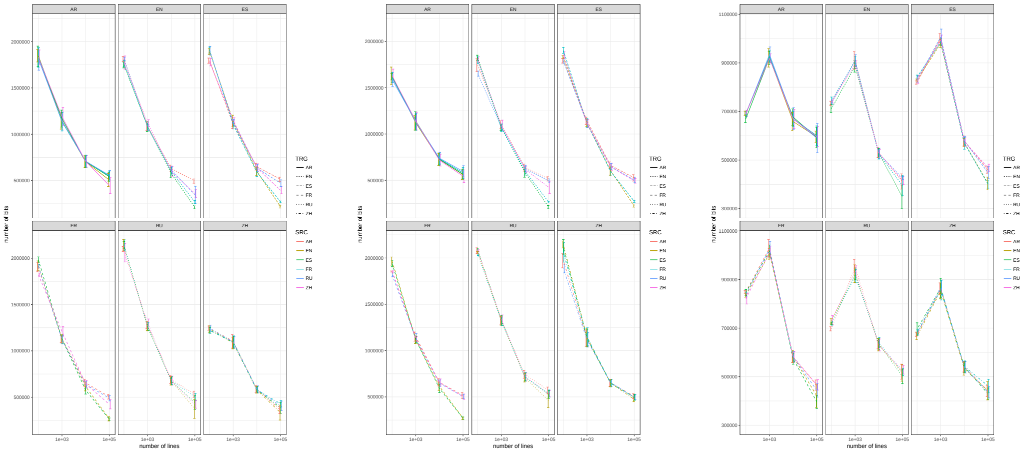
(a) CHAR by target (b) BYTE by target (c) WORD by target

Figure 7: 1-layer models (sorted in 6 facets by target language and with error bars)



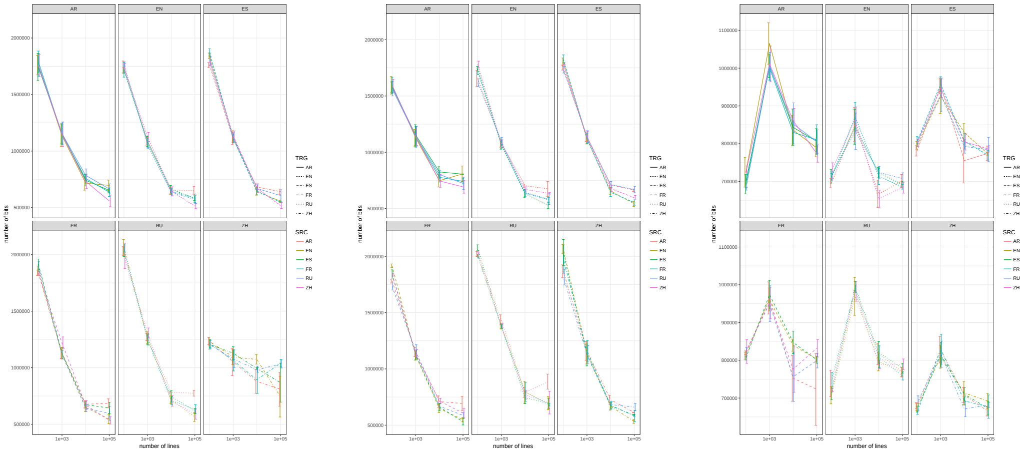
(a) CHAR by target (b) BYTE by target (c) WORD by target

Figure 8: 2-layer models (sorted in 6 facets by target language and with error bars)



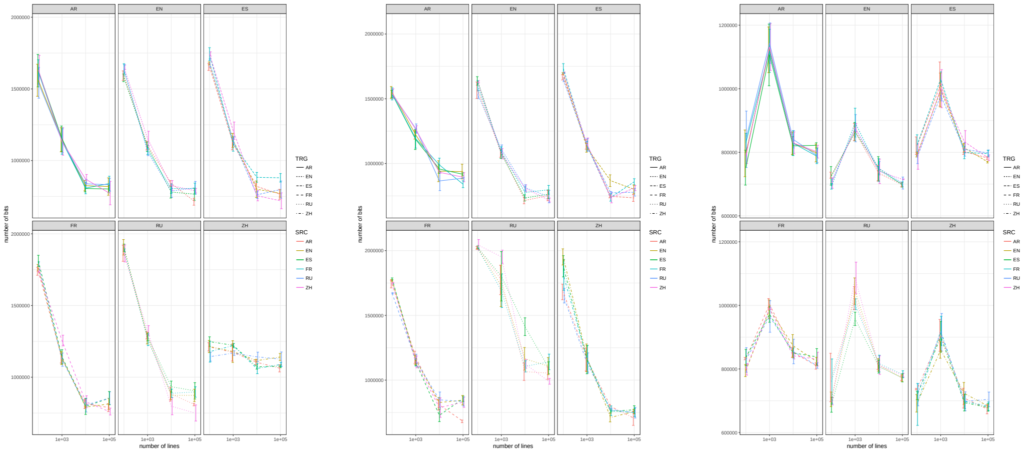
(a) CHAR by target (b) BYTE by target (c) WORD by target

Figure 9: 3-layer models (sorted in 6 facets by target language and with error bars)



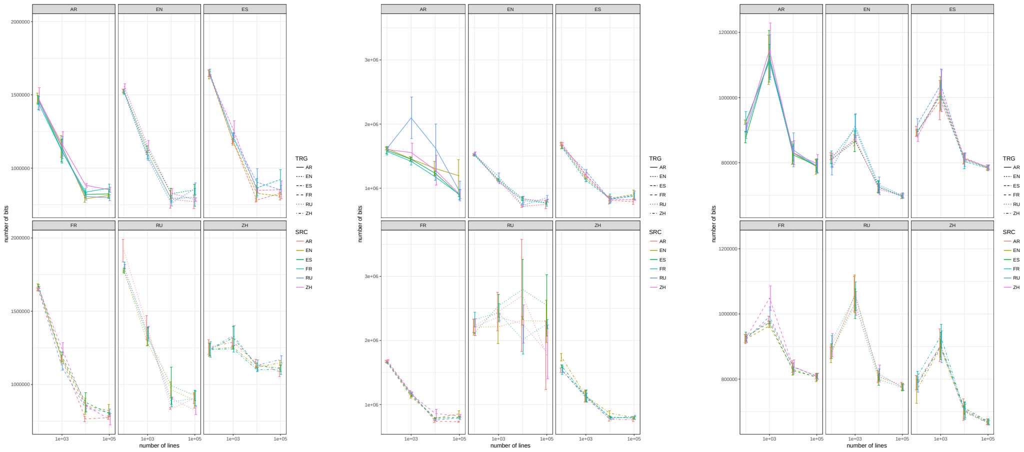
(a) CHAR by target (b) BYTE by target (c) WORD by target

Figure 10: 4-layer models (sorted in 6 facets by target language and with error bars)



(a) CHAR by target (b) BYTE by target (c) WORD by target

Figure 11: 5-layer models (sorted in 6 facets by target language and with error bars)



(a) CHAR by target (b) BYTE by target (c) WORD by target

Figure 12: 6-layer models (sorted in 6 facets by target language and with error bars)

Table 1: Number of language pairs out of 15 possible ones with statistically significant differences, with respective p-values.

	CHAR		BYTE		WORD			CHAR		BYTE		WORD	
1 layers	src	trg	src	trg	src	trg	4 layers	src	trg	src	trg	src	trg
$p = 0.05$	0	**1**	0	0	0	0	$p = 0.05$	0	0	0	0	0	**9**
$p = 0.01$	0	**1**	0	0	0	0	$p = 0.01$	0	0	0	0	0	**8**
$p = 0.001$	0	0	0	0	0	0	$p = 0.001$	0	0	0	0	0	**8**
2 layers	src	trg	src	trg	src	trg	5 layers	src	trg	src	trg	src	trg
$p = 0.05$	0	0	0	0	0	0	$p = 0.05$	0	**1**	0	**5**	0	**10**
$p = 0.01$	0	0	0	0	0	0	$p = 0.01$	0	0	0	**5**	0	**9**
$p = 0.001$	0	0	0	0	0	0	$p = 0.001$	0	0	0	**5**	0	**7**
3 layers	src	trg	src	trg	src	trg	6 layers	src	trg	src	trg	src	trg
$p = 0.05$	0	0	0	0	0	**1**	$p = 0.05$	0	**2**	0	**9**	0	**8**
$p = 0.01$	0	0	0	0	0	0	$p = 0.01$	0	**1**	0	**8**	0	**8**
$p = 0.001$	0	0	0	0	0	0	$p = 0.001$	0	0	0	**7**	0	**8**

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

Ada Wan. Representation and bias in multilingual NLP: Insights from controlled experiments on conditional language modeling, 2021. URL `https://openreview.net/forum?id=dKwmCtp6YI`.

Ada Wan. Fairness in representation for multilingual NLP: Insights from controlled experiments on conditional language modeling. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=-llS6TiOew`.

Microscopes and telescopes: Trading in black boxes for a lens with multitexts, network depths, and statistical comparisons

Draft version 0.01 uploaded 20240425 (work in progress)